

QCDOC Project Hardware Status

2004 BNL All-Hands Meeting

March 26, 2004

Norman H. Christ

OUTLINE

- Project overview.
- Progress and schedule.
- Performance and cost.

QCDOC PROJECT

- Physics: Lattice QCD is driven by a synergy between exponentially increasing computer resources and dramatically improved algorithms.
- Architecture: Large gains possible from optimized design.
 - Space-time homogeneity supports easy parallelization and a mesh network.
 - System-on-a-chip technology permits a highly scalable and cost-effective design:
 - * Entire node (including interconnect logic) on a single chip.
 - * The only extra components:
 - Serial nearest-neighbor wires.
 - Commercial Ethernet tree for booting, diagnostics and I/O.
 - Low power, compact design.
- Goal: \$1/sustained Mflops.

QCDOC COLLABORATION

Columbia (DOE):

Norman Christ
Saul Cohen
Calin Cristian
Zhihua Dong
Changhoan Kim
Ludmilia Levkova
Xiaodong Liao
Guofeng Liu
Robert Mawhinney
Azusa Yamaguchi

UKQCD (PPARC):

Peter Boyle
Mike Clark
Balint Joo

RBRC (RIKEN):

Shigemi Ohta (KEK)
Tilo Wettig (Yale)

IBM:

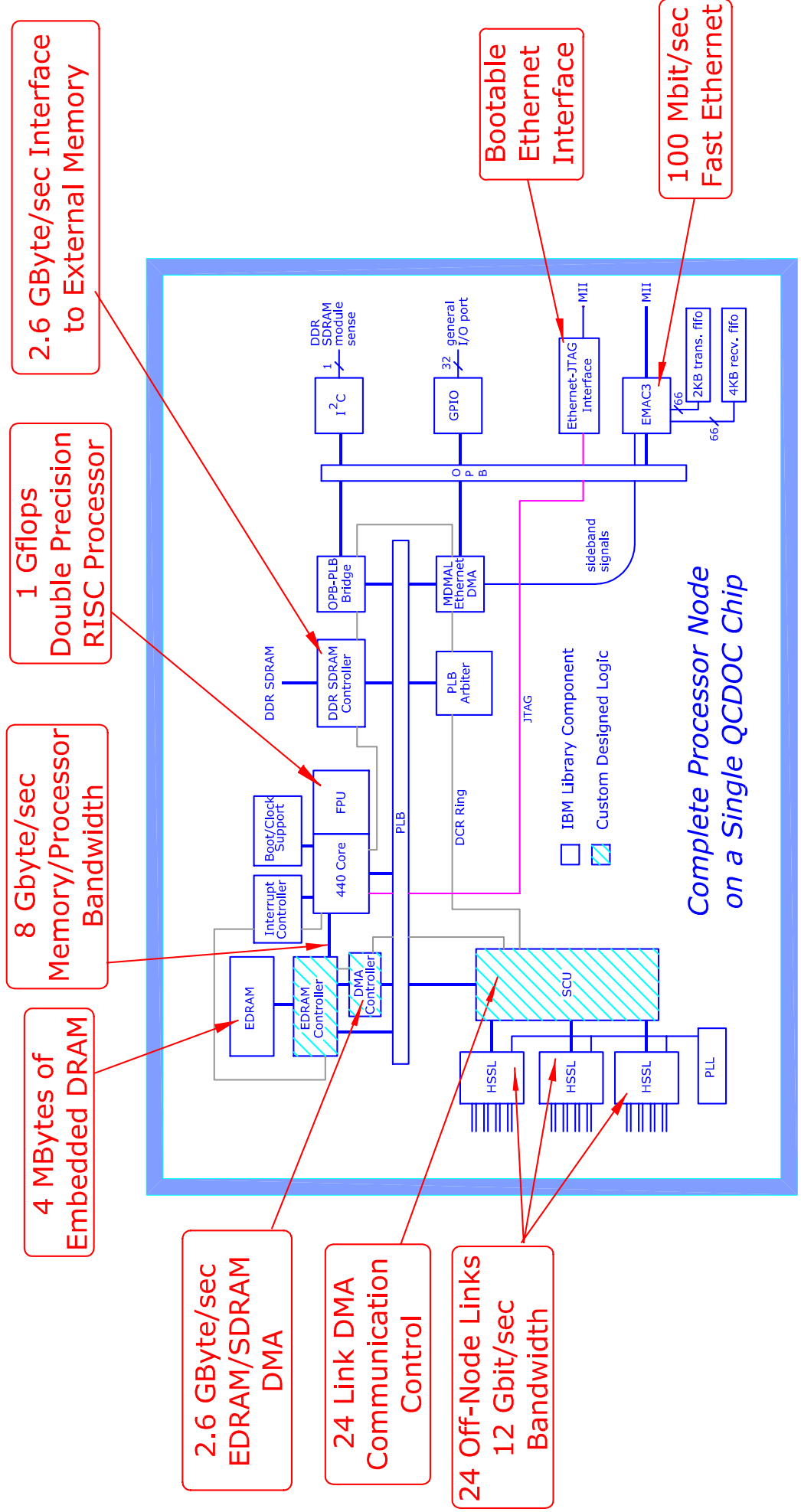
Dong Chen
Alan Gara
Design groups:
Yorktown Heights, NY;
Rochester, MN; Raleigh, NC

BNL (DOE):

Robert Bennett
Chulwoo Jung
Kostya Petrov
Dave Stampf

DESIGN

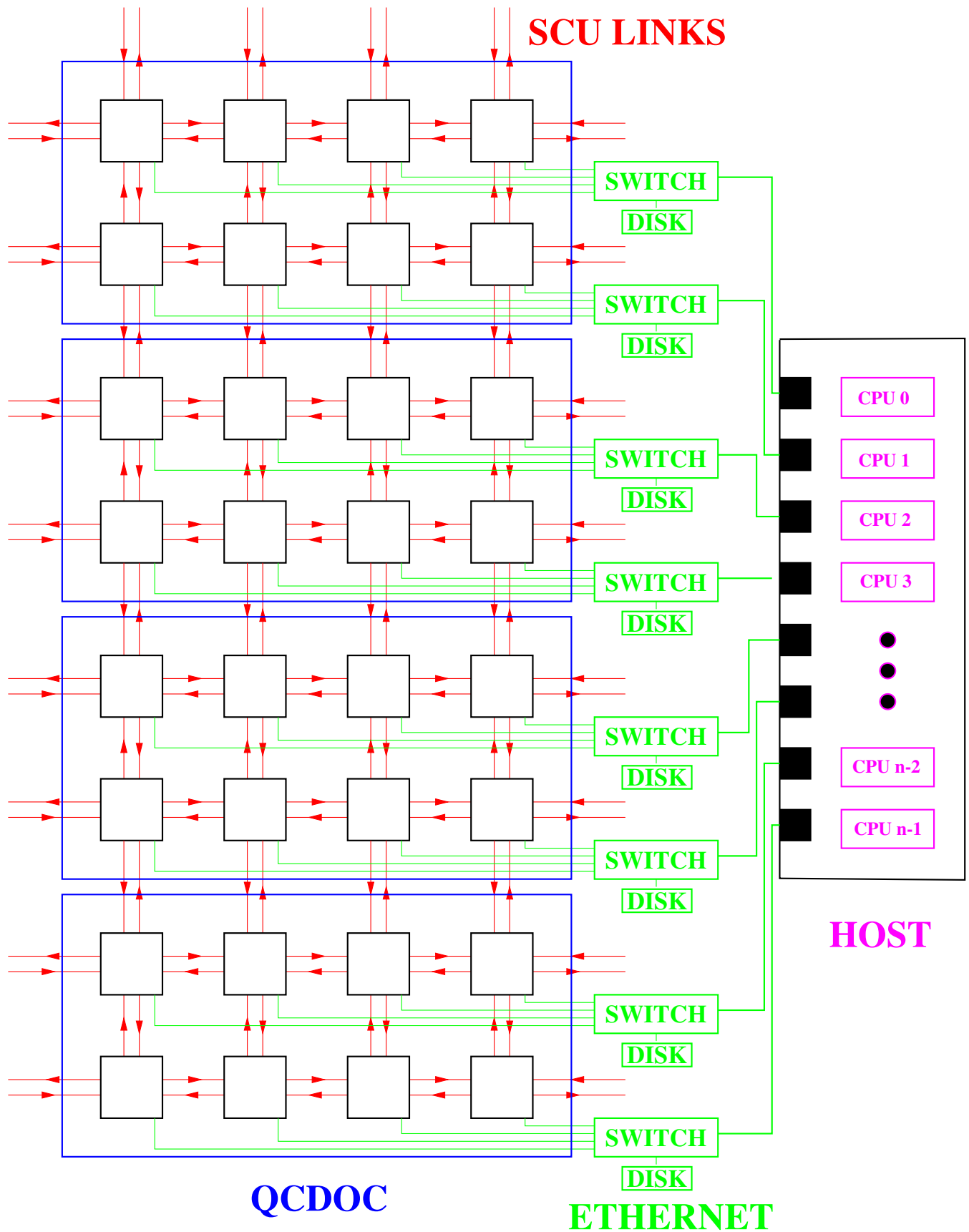
- IBM-fabricated, single-chip node.
[50 million transistors, 5-6 Watt, 1.3cm×1.3cm die]
- PowerPC 32-bit processor
 - 1 Gflops, 64-bit IEEE FPU.
 - Memory management.
 - GNU and XLC compilers.
- 4 Mbyte on-chip memory and up to 2.0 Gbyte/node on DIMM card.
- 6-dim communications network:
 - Efficient for small packet sizes, $\approx 500\text{ns}$ latency.
 - Global sum/broadcast functionality.
 - Minimal processor overhead.
 - Lower dimensional machine partitions.
- 100 Mbit/sec, Fast Ethernet
 - JTAG/Ethernet boot hardware.
 - Host-node OS communication.
 - Disk I/O.
 - RISCWatch debugger.
- ≈ 10 Watt, 15 in³ per node.



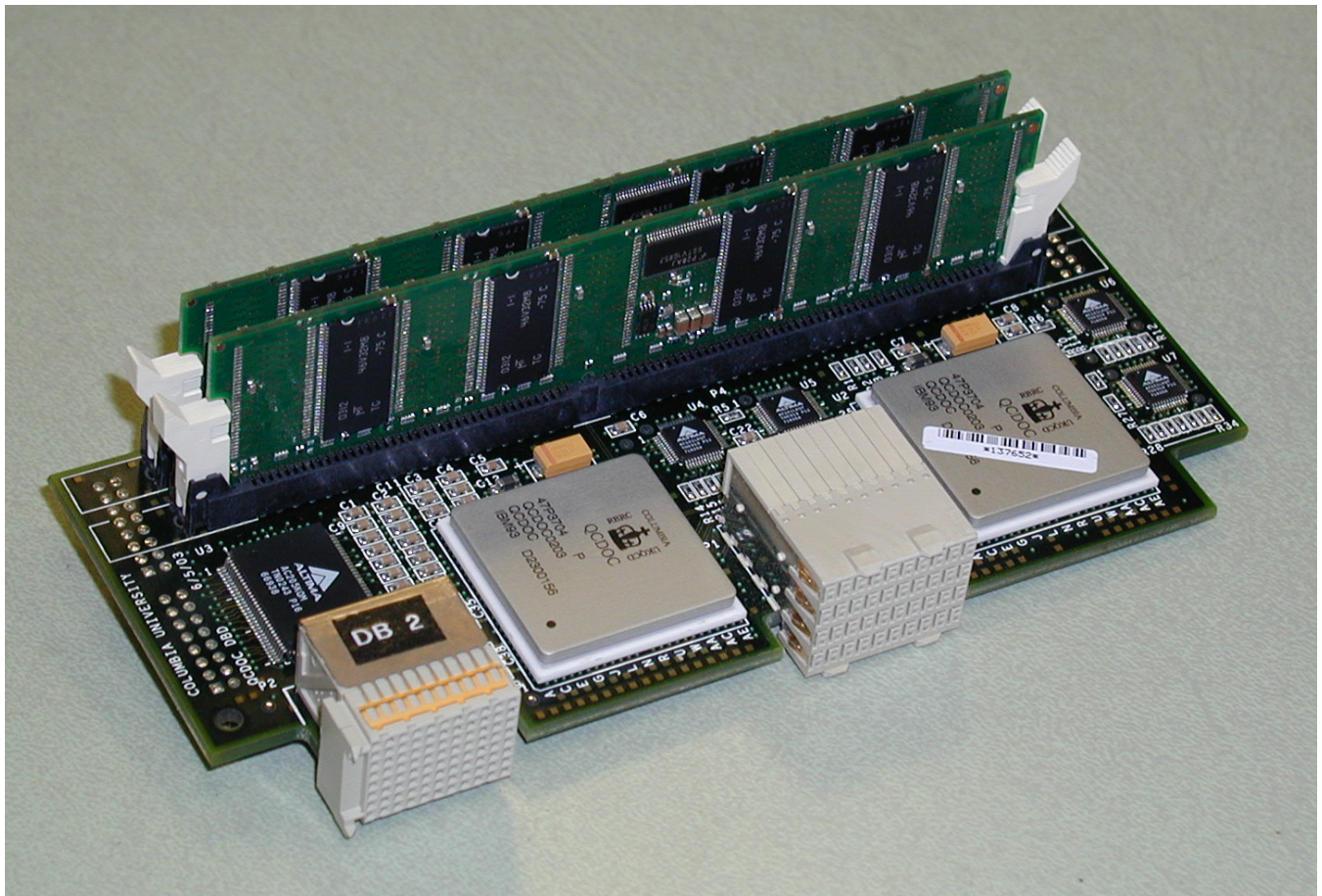
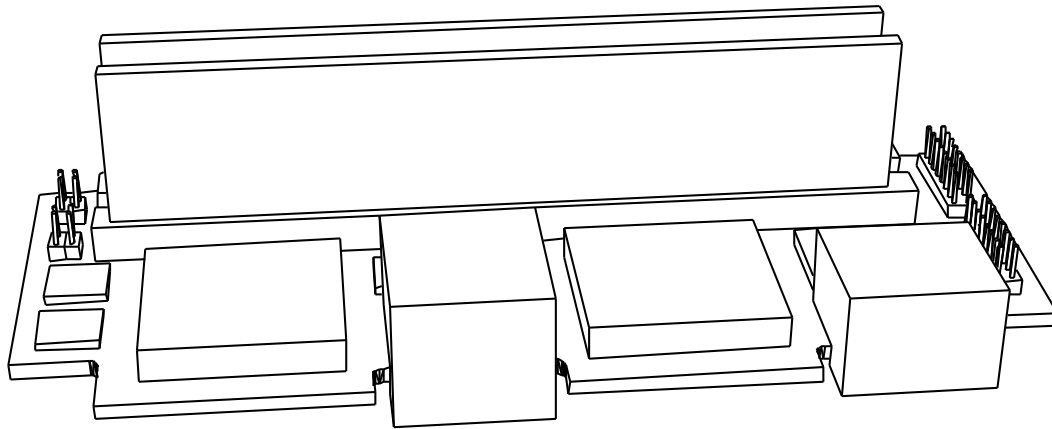
RELIABILITY

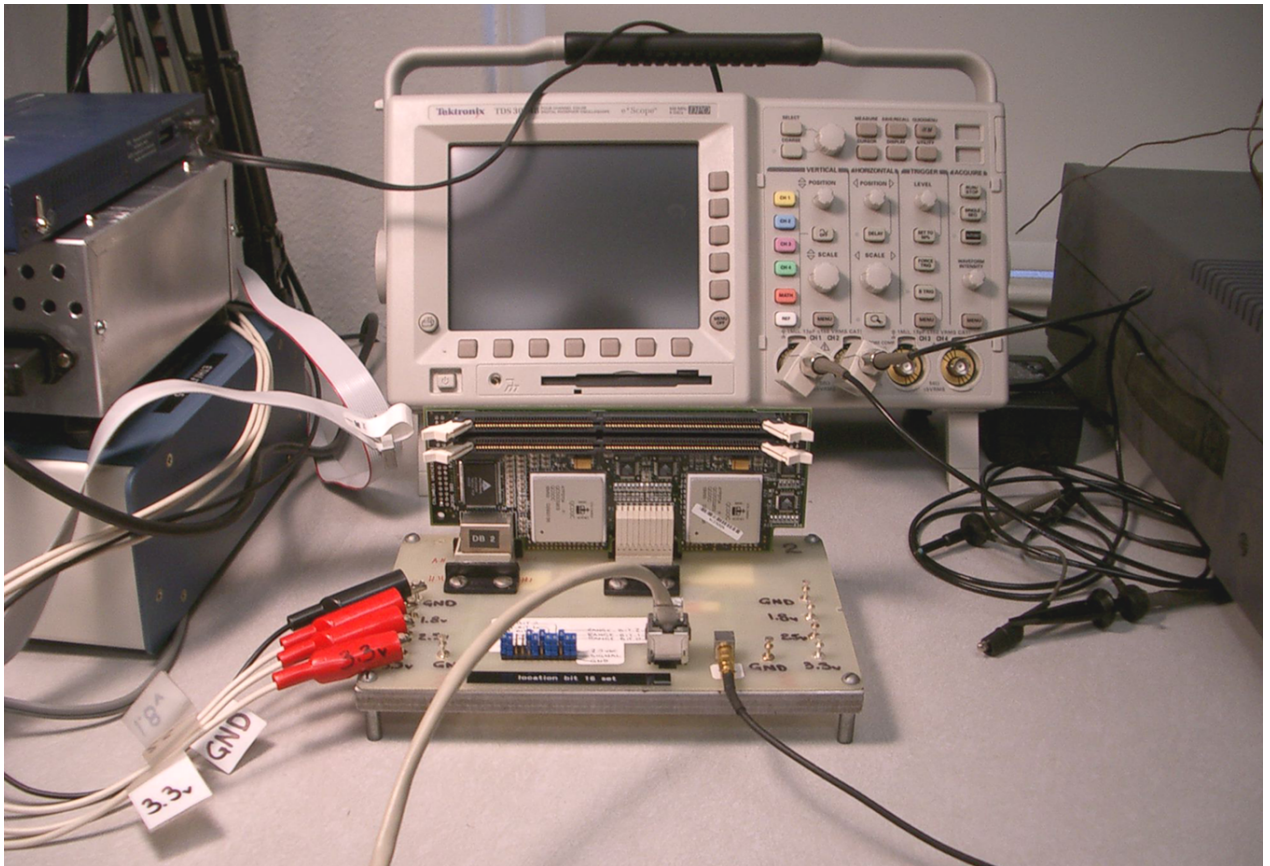
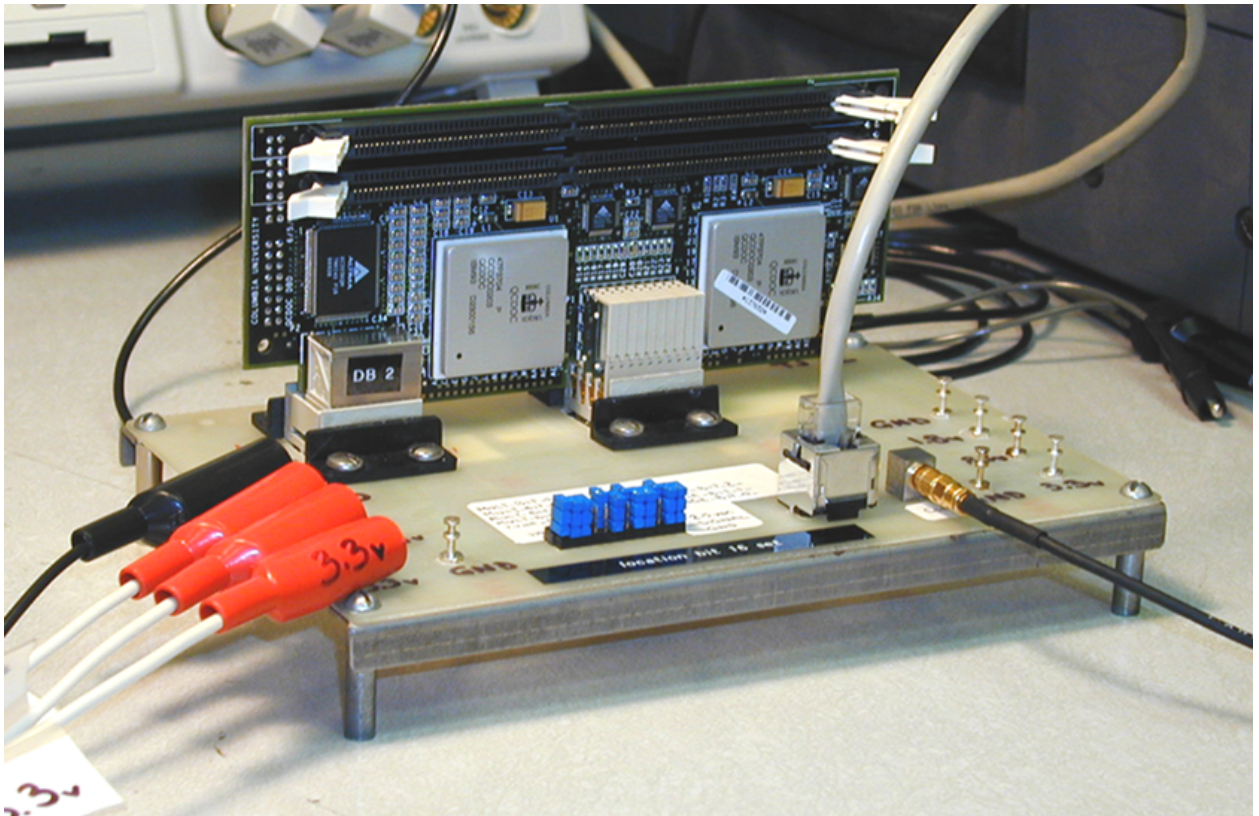
- ECC on external DIMM and EDRAM.
- Automatic recovery from single-bit communications errors.
- Running check sum on both ends of each serial channel.
- Number of components similar to QCDSP: 1-2 failures/week on 10K node machine.
- Soft error rate estimated at $< 1/\text{week}$ on 10K nodes (low- α lead in solder balls).

MACHINE OVERVIEW

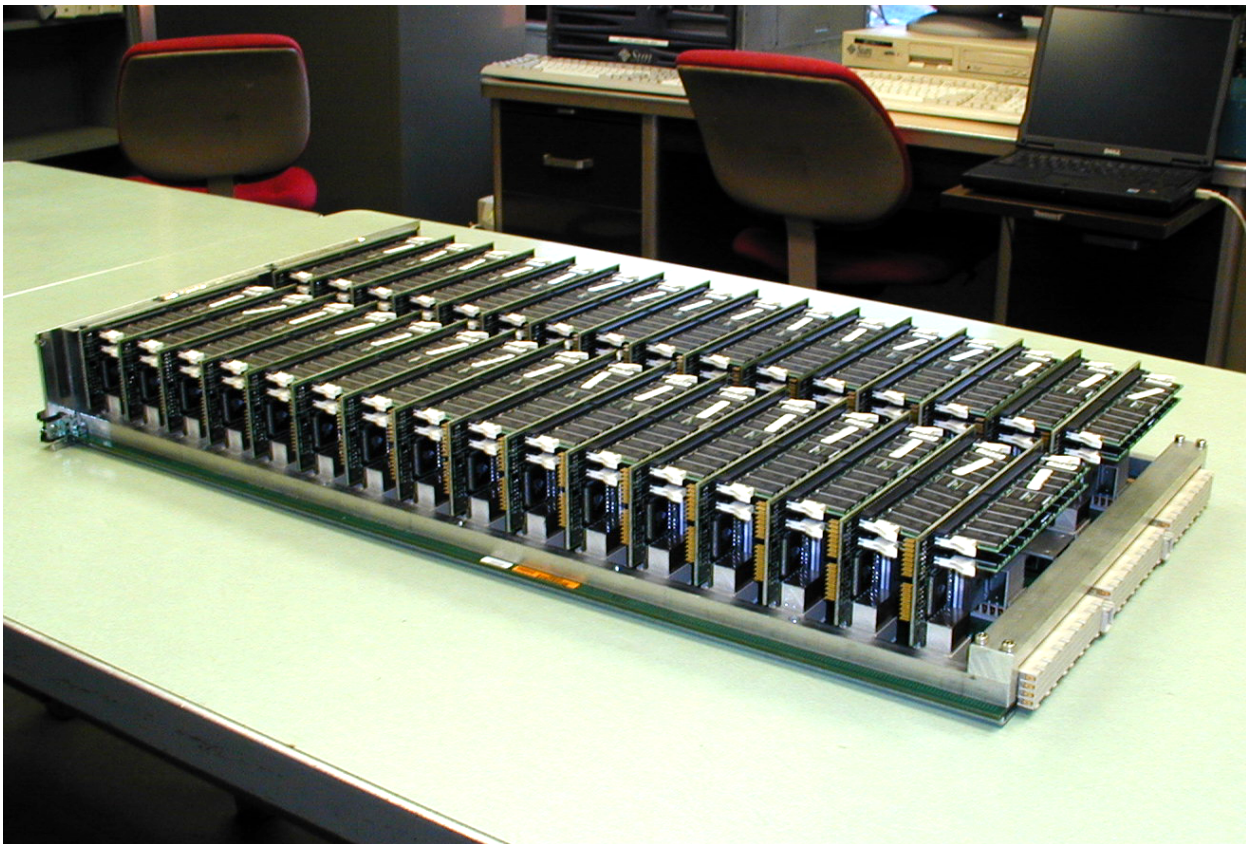
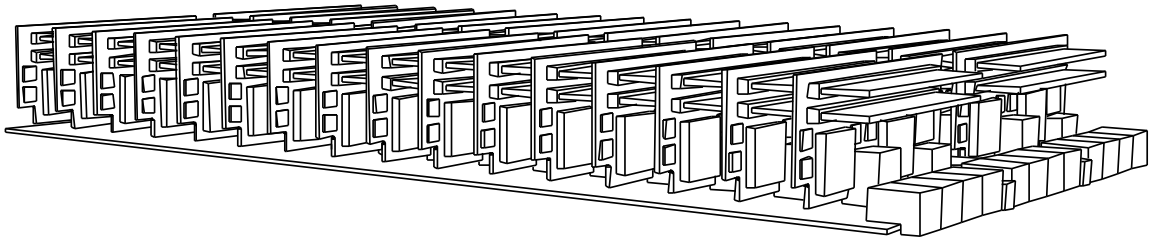


DAUGHTER BOARDS

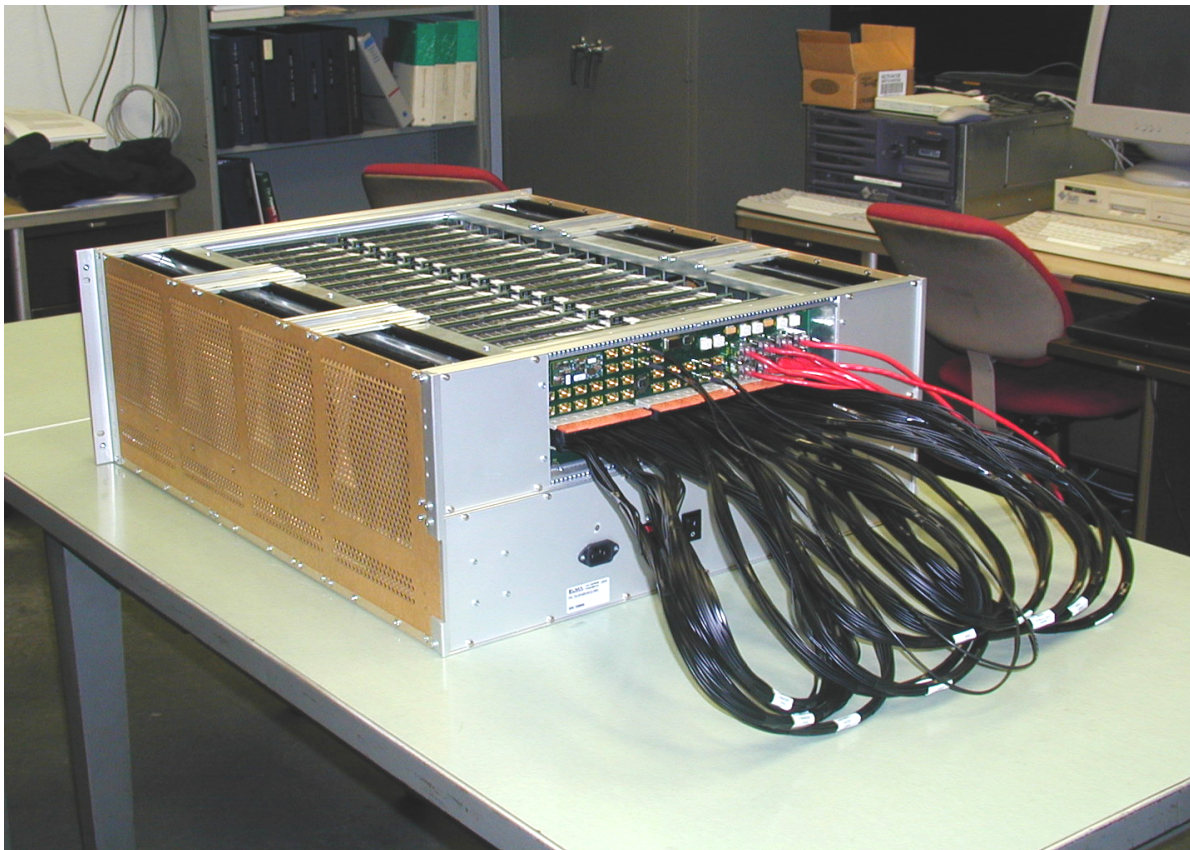
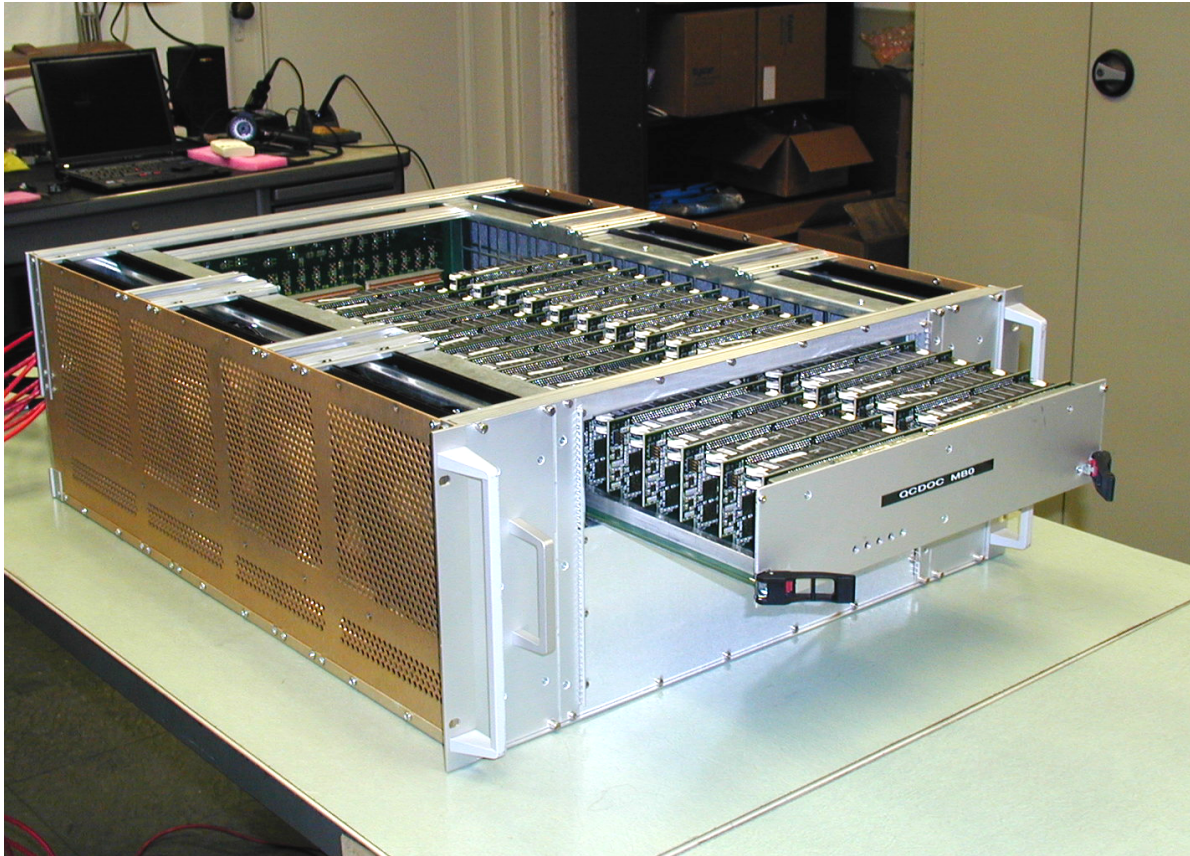




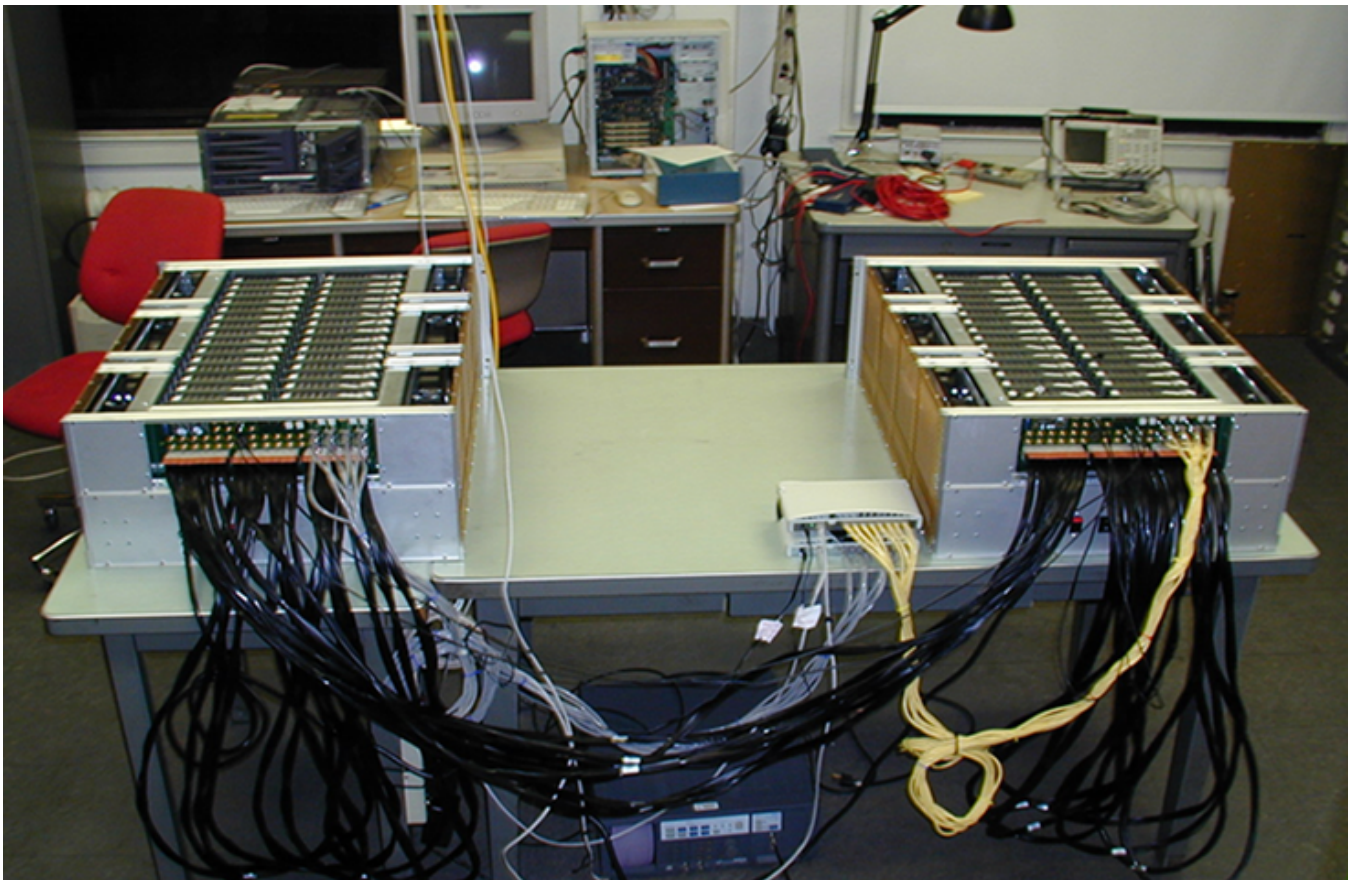
MOTHER BOARDS



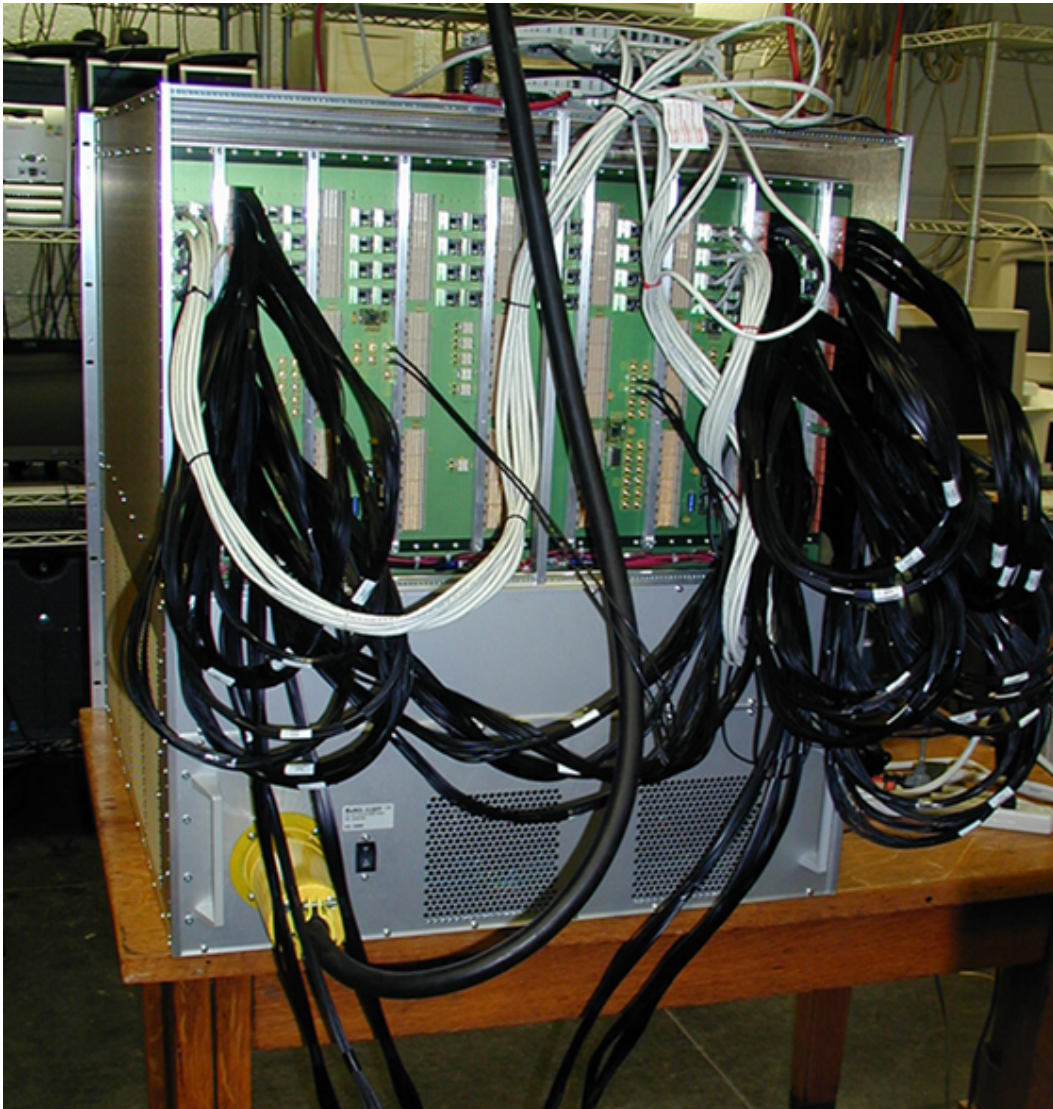
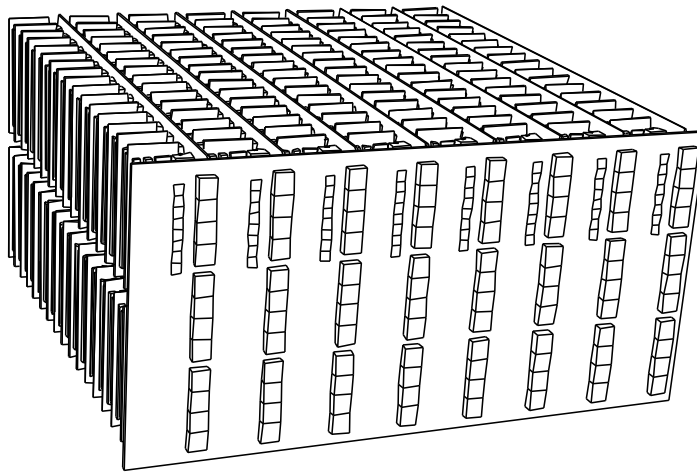
Single Motherboard Cabinet



128-node Machine



8 MOTHER BOARD BACKPLANE

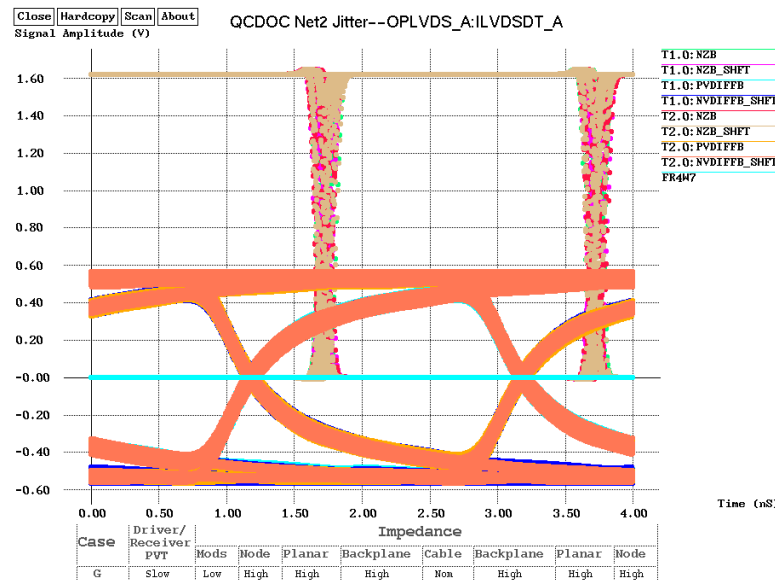


8 MOTHER BOARD CABINET

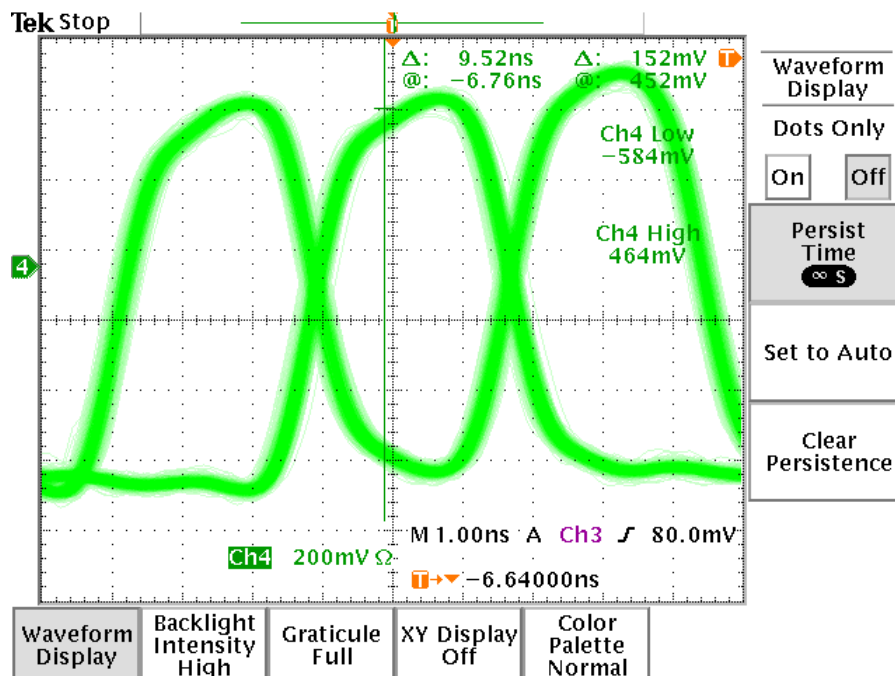


Serial communications simulated and tested

- Simulation:



- Real 6 meter cable:



HARDWARE STATUS

- ASIC tape-out April 8, 2003
- First five chips: June 5, 2003.
- First five daughter cards June 27, 2003
- Basic functionality verified July 9, 2003.
- First three mother boards Sept. 2003.
- Full motherboard functioning Nov. 3, 2003.
- Two motherboards functioning Nov. 18, 2003.
- ASIC sign-off, Nov. 2003.
- Final daughter board sign-off, Mar. 1, 2004.
- Backplane sign-off, Mar. 23, 2004.

COMPLETION SCHEDULE

- Final mother board sign-off, April 7, 2004.
- 384-node machine, April 9, 2004.
- First water-cooled cabinet, May 17, 2004.
- Two 2048-node machines May 31, 2004.
- Two 10,240-node machines, August 31, 2004
[RIKEN, UKQCD].
- Third 10,240-node machine, Oct 31, 2004
[U.S. LQCD Collaboration].

PERFORMANCE and COST

- Initial target:

- Processor frequency: 500 MHz.
- Total cost per node: \$500.
- Price/performance:

$$\frac{\$500}{(2 \text{ flops/cycle}) \cdot 500\text{MHz} \cdot 0.50 \text{ eff.}} = \frac{\$1}{\text{Mflops}}$$

- Present status:

- Processor frequency: 450 MHz.
- Total cost per node: \$400.
- Price/performance:

$$\frac{\$400}{(2 \text{ flops/cycle}) \cdot 450\text{MHz} \cdot 0.50 \text{ eff.}} = \frac{\$0.89}{\text{Mflops}}$$

- Goal of \$1/Mflops should be achieved.
(Exceeded?)